Biophysical Journal Volume 99 November 2010 3029–3037

Characterizing Protein Energy Landscape by Self-Learning Multiscale Simulations: Application to a Designed β-Hairpin

Wenfei Li[†‡] and Shoji Takada[†*]
[†]Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto, Japan, and CREST, Japan Science and Technology Agency, Kawaguchi, Japan; and [‡]Department of Physics, Nanjing University, Nanjing, China

ABSTRACT Characterizing the energy landscape of proteins at atomic resolution is still a very challenging problem, since it simultaneously requires high accuracy in estimating specific interactions and high efficiency in conformational sampling. Here, for these two requirements to meet, we extended the self-learning multiscale simulation (SLMS) method developed recently and applied it to the designed β-hairpin CLN025. The SLMS integrates all-atom and coarse-grained (CG) models in an iterative way such that the conformational sampling is performed by the CG model, the AA energy is used to calibrate the energy landscape, and the CG model is improved by the calibrated energy landscape. We extended the SLMS in two aspects, use of the energy decomposition for self-learning of the CG potential and a two-bead/residue CG model. The results show that the self-learning greatly improved the CG potential, and with the derived CG potential, the β-hairpin CLN025 robustly folded to the native structure. The self-learning iteration progressively enhanced the context dependence in the CG potential and increased the energy gap between the native and the denatured states of the CG model, leading to a funnel-like energy landscape. By using the SLMS method, without prior knowledge of the native structure but with the help of the AA energy, we can obtain a tailor-made CG potential specific to the target protein. The method can be useful for de novo structure prediction as well.

INTRODUCTIONS

At this time, a major problem in the study of protein folding is to characterize the complete energy landscape of proteins at atomic resolution: To what extent is the energy landscape funnelled (1)? Which of the physical interactions contributes to the funnel-like shape? How does the remaining frustration affect the folding? Characterizing the energy landscape at high resolution is challenging, because it requires accurate estimate of atomic interactions, especially near the native structure, along with efficient sampling of a broad range of conformational space, which is responsible for the denatured state. High-resolution description calls for an all-atom (AA) protein model, where a gold standard is the molecular dynamics (MD) simulation with the molecular mechanics force field (2–4). In practice, however, due to the tremendous number of degrees of freedom and the rugged energy landscape, the timescale reachable at present by the AA simulations is below the folding timescales of most proteins, which makes the de novo folding simulations applicable only to very small proteins. To meet the second requirement, i.e., efficient sampling, coarse-grained (CG) models have been widely used in folding and functional studies (5–11). In CG models, each residue is simplified to one or a few CG beads. This greatly reduces the degrees of freedom and the ruggedness of the landscape, thus speeding up the conformational sampling. It appears that such CG simulation is meaningful only when the CG potential used can capture some aspects of the overall structure and dynamics of the proteins. However, modeling the realistic interactions effectively by a reduced protein representation is not straightforward. It comes as no surprise that very different methods have been developed and reported in the literature (5–20).

One common problem in generic CG models is that by reducing the side-chain representation, CG models unavoidably lose specificity in side-chain interactions, which tends to destabilize the native state relative to the denatured state. To solve this problem, we put forward the tailor-made CG model, i.e., the CG potential tuned to a specific target protein. The specific CG model can have context-dependent interactions by which we can take into account specific interactions near the native structure. Here, we address how the specific and tailor-made CG potential can be derived by so-called multiscale simulation methods, which have been attracting much attention in recent years (16–32).

In a major class of the multiscale protocols, which we call AA→CG protocol, the CG force field is derived based on the force/energy information or the sampled conformational distributions of the AA simulations by using a certain learning method, e.g., force matching (18), iterative Boltzmann inversion (IBI) (27), fluctuation matching (17,22), etc. By using such a bottom-up strategy, it is possible to maintain the interaction specificity of the AA force field during the coarse-graining. It was demonstrated that the CG potential derived based on the above multiscale protocol can capture a number of structural and dynamic features for

Submitted April 27, 2010, and accepted for publication August 18, 2010.
*Correspondence: takada@biophys.kyoto-u.ac.jp
Editor: Gregory A. Voth.

0006-3495/10/11/3029/9 $2.00
doi: 10.1016/j.bpj.2010.08.041

some biomolecules (16–22,33). However, by definition, accuracy of the derived CG potentials by the AA→CG protocol is totally encoded by the parent AA simulations used to derive them. In principle, the derived CG potential is accurate only in the conformational space covered by the parent AA simulations which is often quite narrow due to the low sampling efficiency. Such a feature makes the application scope of the above multiscale protocol limited.

To overcome this difficulty, in our previous works, we developed the self-learning multiscale (SLMS) method in which CG simulation and AA simulation are tightly coupled together and the CG potential is derived in an iterative way based on the previously sampled CG conformations and their corresponding AA energies (29,30). In contrast to the AA→CG protocol, in the SLMS method, conformational sampling is always controlled by the CG simulations and therefore is more efficient. The AA energies are embedded simultaneously into the CG potential to shape up the CG energy landscape. By such a two-way integration of the AA and CG models, the derived CG potential can be accurate in a much wider conformational space and therefore can be directly used for simulations in CG level. Alternatively, the improved CG potential can be used as the input of other two-way coupling multiscale methods (34–36). For example, in the resolution exchange methodology developed by Zuckerman and co-workers (34,35), the AA MD is coupled to a certain CG MD via trials of conformational exchange. Since the CG MD usually covers a much wider range of conformational space, exchange of conformations can significantly enhance the AA MD sampling. It appears that reasonably accurate CG potential is the prerequisite for efficient exchange of conformations, and the SLMS simulations can be an effective way to prepare the required CG potential (29).

In this work, we extended the SLMS method in two ways and applied it to the folding of a designed β-hairpin. In our previous work, each residue was coarse-grained to one bead, and, the self-learning of the CG potential was accomplished by the IBI (26,27). As noticed also by other people (37), in deriving the pairwise CG potential with Boltzmann inversion, the many-body distribution functions need to be decomposed into two-body distribution functions, which can diminish the accuracy of the derived CG potentials since the pairwise CG potential, optimized only by fitting the two-body distributions, cannot always reproduce many-body distributions. Similar discussions can also be found in Harmandaris et al. (25). Here, instead of using the IBI, we used energy decomposition as the learning method by which to derive the nonlocal part of the CG potential (30,38). Energy decomposition extracts information about the CG potential from AA energies at the interaction level, thus eliminating the problem encountered using IBI. In addition, the one-bead/residue approximation implies that the pairwise interactions between the interacting residues are isotropic and therefore cannot reflect the environment variations around the CG beads. However, in reality, the interactions between the interacting residues can be highly related to their relative orientations. In this work, to improve the interaction specificity of the derived CG potential, each residue was coarse-grained to two beads, namely, the backbone bead and side-chain bead, which ensured more realistic representation of the relative orientations of residues.

After illustrating the SLMS method by applying it to a pentaalanine peptide, we used it to optimize the CG energy function for the designed β-hairpin CLN025 (39). For this β-hairpin, sufficient sampling can be achieved even by AA simulations, and therefore, it can be used to evaluate the SLMS method unambiguously. Test results showed that the derived CG potential can successfully fold CLN025. In particular, due to iterative self-learning, the energy landscape acquired a funnel-like behavior, with the most stable state being highly biased to the native state. Such a feature is essential for the successful recognition of the folded structure, suggesting that the SLMS method is capable of capturing the overall features of the realistic energy landscape in deriving the CG potential and therefore can be used for the purposes of protein folding and de novo structure prediction.

METHODS

Structure mapping between CG and AA models

Two-bead/residue coarse-graining was used in the structure mapping from AA to CG models. For each residue, one bead located at the C_α position represents the backbone atoms and another bead located at the center of mass of the side chain represents all the side-chain atoms (glycine was represented by one bead). For the reverse structure mapping, from CG to AA models, we used the software BBQ (40) and SCWRL (41), which reconstruct the backbone and side-chain atoms, respectively. The reconstructed AA structures were then minimized by the AA force field. During the minimization, the coordinates of the C_α atom and the heavy atom closest to the side-chain center of mass were restrained to the positions of the corresponding CG beads in the CG structure. As the minimization proceeded, the weights of the van der Waals and Coulombic interactions were gradually increased from 0 to 1. The minimized structures were further heated to 600 K to remove possible bad interactions, with the C_α atoms being restrained to their original positions.

The above reconstruction scheme gives only one AA structure for each CG structure. However, in practice, each CG structure may have a number of corresponding AA structures with different relative weights. As an effort to remedy this problem, we conducted short AA MD simulations starting from the above reconstructed AA structure at the target temperature (250 K for pentaalanine and 300 K for β-hairpin CLN025). All the sampled structures in the equilibrium MD were considered to be the corresponding AA representations and were used in the subsequent analysis. Even so, the above procedure does not capture all the relevant AA structures, which can result in some errors, as discussed later.

The schematic diagram of the structure mapping is shown in Fig. 1, with the backbone beads labeled b and the side-chain beads labeled s. The CG potential, including the local part and the nonlocal part, is given by

$$V(R) = \sum_{bb} V_b(\mathbf{R}^b_m, \mathbf{R}^b_{m+1}) + \sum_{bs} V_b(\mathbf{R}^b_m, \mathbf{R}^s_m) + \sum_{bbb} V_\theta(\mathbf{R}^b_{m-1}, \mathbf{R}^b_m, \mathbf{R}^b_{m+1}) + \sum_{bbs} V_\theta(\mathbf{R}^b_{m-1}, \mathbf{R}^b_m, \mathbf{R}^{bs}_m) + \sum_{sbb} V_\theta(\mathbf{R}^s_m, \mathbf{R}^b_m, \mathbf{R}^b_{m+1}) + \sum_{bbbb} V_\phi(\mathbf{R}^b_{m-2}, \mathbf{R}^b_{m-1}, \mathbf{R}^b_m, \mathbf{R}^b_{m+1}) + \sum_{bbbs} V_\phi(\mathbf{R}^b_{m-2}, \mathbf{R}^b_{m-1}, \mathbf{R}^b_m, \mathbf{R}^s_m) + \sum_{sbbb} V_\phi(\mathbf{R}^s_m, \mathbf{R}^b_m, \mathbf{R}^b_{m+1}, \mathbf{R}^b_{m+2}) + \sum_{nloc} V_{nl}(\mathbf{R}^{b,s}_m, \mathbf{R}^{b,s}_n), \quad (1)$$

where $\mathbf{R}^b_m$ and $\mathbf{R}^s_m$ represent the positions of the backbone and side-chain beads, respectively, of the mth residue. The V_b, V_θ, V_ϕ, and V_{nl} are the bond length, bond angle, dihedral angle, and nonlocal interaction terms, respectively. The summation index bb (bs) runs over all the bonds formed by the adjacent CG backbone beads (backbone and side-chain beads). In a similar way, we can define the summation indices for other terms. The nonlocal interaction involves all the CG bead pairs separated by N residues in sequence with $N \geq 4$, 3, and 3 for the b-b, b-s, and s-s pairs, respectively. We used the data table to represent the interactions, except in the case of the bond-length term, for which a harmonic energy function was used. During the MD simulations, the interactions in the data table were smoothed and interpolated by cubic spline function to give the energies and forces. We emphasize that, even though some of the amino acids are chemically identical, their CG beads represent different types of interaction sites taking into account context dependence.

Self-learning multiscale method

In the SLMS method, the CG potential is derived in an iterative way based on the conformations sampled by the CG model and the energies estimated by the AA force field (see Li and Takada (29) for details). In this work, the local part of the CG potential was derived based on the IBI learning method (26). For the nonlocal part, we used energy decomposition (38) as the learning method. Specifically, we started with an arbitrarily chosen CG potential. After the CG simulations, we mapped all the sampled CG structures to the AA representations and computed their corresponding AA energies, simultaneously using the energy decomposition utility implemented in AMBER software (42) to decompose the AA energy of each structure into the interactions of atomic pairs according to the method developed in Gohlke et al. (38). The interactions between the nonlocal interacting CG beads in each structure were then given by $U^k_{CG,IJ} = \sum_{i\in I}\sum_{j\in J} u^k_{AA,ij}$, where the $u^k_{AA,ij}$ is the decomposed AA interaction between atoms i and j in the kth reconstructed AA structure. The $U^k_{CG,IJ}$ is the CG interaction between the corresponding beads I and J. More details on the energy decomposition can be found in the above-mentioned studies (38,42). We can then define the nonlocal part of the CG interaction, $V_{nl}(R_{IJ})$, by reweighted averaging (43):

$$V_{nl}(R_{IJ}) = \frac{\sum_k \delta(R^k_{IJ} - R_{IJ}) U^k_{CG,IJ} \exp(\beta_{CG}E_{CG}(\mathbf{X}_k) - \beta_{AA}E_{AA}(\mathbf{X}_k))}{\sum_k \delta(R^k_{IJ} - R_{IJ}) \exp(\beta_{CG}E_{CG}(\mathbf{X}_k) - \beta_{AA}E_{AA}(\mathbf{X}_k))}, \quad (2)$$

with $\beta_{CG} = 1/(k_B T_{CG})$ and $\beta_{AA} = 1/(k_B T_{AA})$. E_{CG} and E_{AA} are the corresponding CG and AA energies, respectively. $\boldsymbol{X}_k$ stands for the coordinate of the kth structure. The $R_{IJ}(R^k_{IJ})$ is the distance between residues I and J (in the kth structure). To derive the local term of the CG potential based on the IBI, we need to calculate the distribution function, $g(q)$, corresponding to the local coordinate q, including the bonds, bond angles, and dihedral angles defined above. The $g(q)$ was calculated based on the structures sampled by the CG MD simulations. To eliminate errors resulting from poor CG potential, these CG structures were reweighted according to their AA energies by

$$g(q) = \frac{\sum_k \delta(q_k - q) \exp(\beta_{CG}E_{CG}(\mathbf{X}_k) - \beta_{AA}E_{AA}(\mathbf{X}_k))}{\sum_k \exp(\beta_{CG}E_{CG}(\mathbf{X}_k) - \beta_{AA}E_{AA}(\mathbf{X}_k))}. \quad (3)$$

From $g(q)$, the initial guess of the local CG interactions, $V_q(q)$, can be estimated by $V_q(q) = -k_B T \ln(g(q)/g_R(q))$, where k_B is the Boltzmann constant, and the $g_R(r)$ is the reference distribution for each local coordinate ($g_R(\theta) = \sin(\theta)$ for the bond angle and $g_R(q) = 1.0$ for other local coordinates). The $V_q(q)$ was further adjusted adaptively by the standard IBI scheme (26) to reproduce the target distribution function $g(q)$. This local interaction, $V_q(q)$, together with the nonlocal interaction, $V_{nl}(R_{IJ})$, given by Eq. 2, were then used as the CG force field for MD sampling during the next learning iteration step.

During the above procedure, the atomic energies, which are assumed to be accurate, were propagated to the CG potential by reweighting and energy decomposition. The reweighting in each self-learning step corrected the relative weight of the sampled structures, which validates the subsequent calculations of the distribution $V_q(q)$ and the averages of the decomposed energies. The above procedure was repeated iteratively until the derived CG potential converged (in this work, each iteration step is considered to be one self-learning iteration). Such an iteration procedure improves the accuracy of the final CG potential by progressively

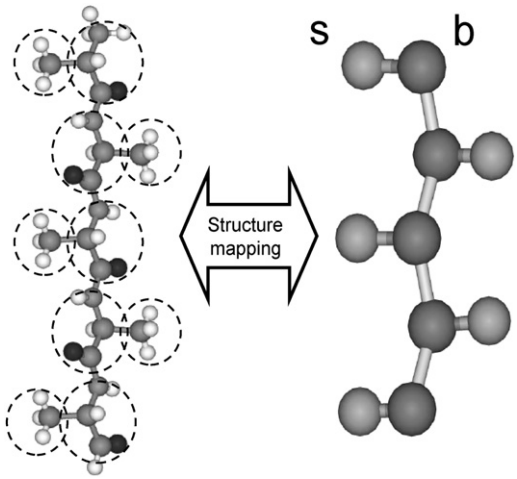

FIGURE 1 Schematic diagram of the structure mapping.

integrating the CG sampling and the AA energy, namely, the advantages of both methods.

During the self-learning, structures with strong pairwise repulsion can hardly be sampled. Therefore, we cannot derive the short-range repulsion part of the nonlocal CG interactions based on the energy decomposition. To overcome this difficulty, we defined a distance cutoff, R_C, with the probability $P(R_{IJ} < R_C) < 0.01$. The CG potential with $R_{IJ} < R_C$ was set as constant: $U_{CG}(R_{IJ} < R_C) = U_{CG}(R_C)$. The excluded volume effect was considered separately by an additional repulsion term, $\varepsilon_{IJ}(\sigma_{IJ}/R_{IJ})^6$ with $\sigma_{IJ} = (\sigma_I + \sigma_J)/2$ and $\varepsilon_{IJ} = (\varepsilon_I\varepsilon_J)^{1/2}$. The parameters σ_I and ε_I were taken from other studies (44,45). We also tested the values of $\varepsilon_I = 4.0$ kcal/mol for the side-chain beads and found that the results did not change significantly.

Simulation details

AA simulations were conducted using the AMBER10 package with the force field ff99SB and GB/SA implicit solvent (42,46,47). For comparison, the standard temperature-replica-exchange MD (T-REMD) (48) with eight replicas was used to obtain a converged sampling for both the pentaalanine peptide (100 ns) and the β-hairpin CLN025 (200 ns). The temperatures of the T-REMD for pentaalanine and CLN025 ranged from 225.0 K to 520.0 K and from 285.0 K to 620.0 K, respectively. The SHAKE algorithm was used to restrain the bond lengths involving hydrogen atoms (49), and a time step of 0.002 ps was used. For CG simulations, the Langevin dynamics with $\gamma = 0.01$ was used. The initial structures of the simulations were prepared by AA MD at 1000 K starting from an extended conformation. To construct the initial CG potential, a short AA MD simulation at 300 K was conducted for 400 ps starting from the above initial structure. Then, Boltzmann inversion and energy decomposition were used to derive the local and nonlocal parts, respectively, of the initial CG potential. At each self-learning iteration step, CG MD of the SLMS simulations was conducted for 5×10^6 MD steps. Since it is not straightforward to theoretically prove the convergence of the current SLMS algorithm, we conducted an independent SLMS simulation for the β-hairpin CLN025 with the α-helical structure being the initial structure. Accordingly, in deriving the initial CG potential, the short AA MD simulation started from the structure given by high-temperature unfolding of the α-helices.

To characterize the folding of the β-hairpin CLN025, the reaction coordinates Q (fraction of native contacts) and root-mean-square deviation (RMSD) from the native structure were used. The weighted histogram analysis method (WHAM) was used to construct the free-energy landscape and calculate the distributions from the T-REMD results for the β-hairpin CLN025 (50). Software VMD was used to illustrate the protein structures (51).

RESULTS AND DISCUSSION

Pentaalanine peptide

As an illustration, we conducted the SLMS simulations for a pentaalanine peptide. Fig. 2 shows the initial CG potentials and the CG potentials derived after two steps of self-learning iterations. Only the interactions involving the backbone beads are shown in Fig. 2. One can see that self-learning can significantly change the CG potential. For example, before the self-learning iterations, the interactions of the bond angles prefer the extended structures ($\theta > 130^\circ$), suggesting that the initial CG potentials derived based on the short AA MD simulations had significant biases to the initial structure. After two steps of self-learning iterations, the helical conformation ($\theta < 100^\circ$) became more

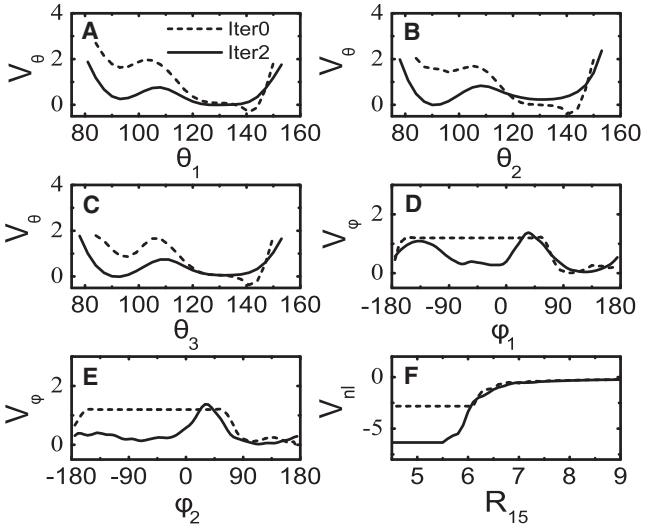

FIGURE 2 CG potentials before self-learning (*dashed lines*) and after two steps of self-learning iterations (*solid lines*).

stable, which is consistent with the empirical observation that alanine has high helical propensity (52). For the dihedral-angle terms, we used a quite unrealistic initial potential in which a constant interaction is assigned to most parts of the dihedral angles, since the initial short parent AA MD cannot sample these regions. After two steps of self-learning iterations, the dihedral-angle potentials were trained and could be assigned for the entire dihedral-angle range due to the improved sampling, demonstrating that the local part of the CG potential can be improved by self-learning iterations, which in turn makes the CG sampling cover a wider range of the relevant conformational space. Fig. 2 *F* shows the nonlocal potential between the first and the last backbone beads. For clarity, the additional repulsion term is not shown. One can see that with the self-learning iterations, the interaction of the shorter-range part was derived. The self-learning-derived CG potentials can reproduce the conformational distributions of the AA T-REMD more closely compared to the initial CG potentials (Fig. S1 in the Supporting Material). These results suggest that the SLMS method can be useful in deriving CG potentials specific to the target system.

The designed β-hairpin CLN025

The designed β-hairpin CLN025 was considered to be to our knowledge the smallest protein with a known structure to date (39). It was demonstrated that this β-hairpin can robustly fold to its x-ray structure by AA MD simulations (39). Such a feature makes it an ideal system to evaluate the ability of the SLMS method to derive a CG potential that can capture the overall behavior of the funnel-like energy landscape. Fig. 3 *A* shows the x-ray structure of CLN025 and its amino acid sequence. As a control experiment, we first conducted AA simulations using T-REMD

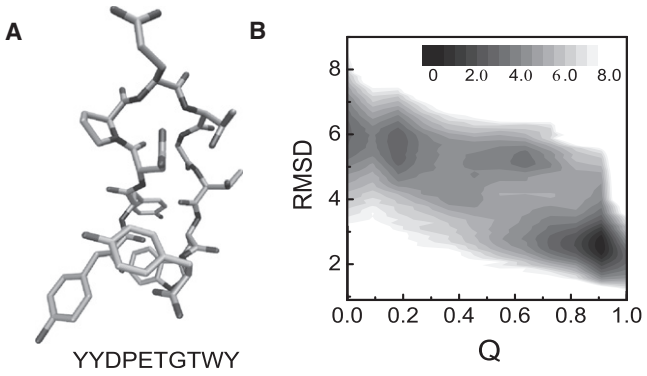

FIGURE 3 (*A*) X-ray structure and sequence of the β-hairpin CLN025. (*B*) Free-energy landscape (k_BT) produced by AA simulations projected onto the conformational space of Q and RMSD.

and computed the folding free-energy landscape. Fig. 3 *B* shows the free-energy landscape projected onto the conformational space formed by Q and RMSD. One can see that with the force field used in this study, the native state of the CLN025 has the lowest free energy and therefore is the most stable state, suggesting that the AA force field is reasonable for the folding of this β-hairpin. Such results rationalize the use of this AA force field to derive the CG potential based on the SLMS method in this work.

By using the SLMS method described in the Methods section, we derived CG potentials and conducted long-time CG simulations (5×10^7 MD steps) for CLN025 at each self-learning iteration stage. All CG potentials were parameterized for simulations at 300 K. Fig. 4 *A* shows

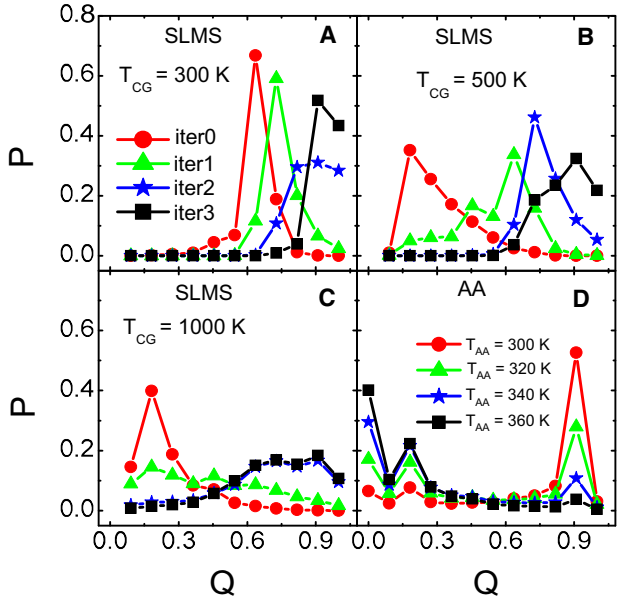

FIGURE 4 (*A–C*) Q distributions before self-learning and after one, two, and three steps of self-learning at $T_{CG} = 300$ K (*A*), $T_{CG} = 500$ K (*B*), and $T_{CG} = 1000$ K (*C*), respectively. (*D*) Results of AA T-REMD simulations at $T_{AA} = 300$ K, 320 K, 340 K, and 360 K.

the probability distributions, $P(Q)$, at $T_{CG} = 300$ K calculated at each self-learning iteration step. One can see that before self-learning (iter0), the conformations with a Q score of ~0.6 represent the most stable state. With the self-learning iterations, the probability distributions are shifted to conformations with a higher Q score. For example, after three steps of self-learning iterations (iter3), the folded state ($Q \sim 1.0$) becomes the most stable state. We also conducted simulations at higher temperatures, e.g., $T_{CG} = 500$ K and $T_{CG} = 1000$ K, the results of which are shown in Fig. 4, *B* and *C*, respectively. Here, we used the CG potentials parameterized for simulations at 300 K, although in principle the temperature-dependent CG potentials can be obtained by reweighting (53). We can see that even at $T_{CG} = 1000$ K, the folded state can still be sampled with a high probability after the self-learning iterations. The free-energy landscape projected onto the Q and RMSD plot at $T_{CG} = 1000$ K also shows movement of the free-energy minimum from the unfolded state to the nativelike state with the self-learning iterations (Fig. S2). These results clearly demonstrate that the CG potentials derived by the SLMS simulations effectively capture the interaction features that are essential for the folding of CLN025. Similar results were obtained for an independent simulation with quite different initial structure and CG potential, though with somewhat different convergence procedures, demonstrating the robustness of this SLMS scheme (Fig. S3). We note that there are quantitative differences among the CG results (Fig. 4, *A–C*) and the AA results (Fig. 4 *D*), which will be discussed later in detail.

To understand the folding ability of the derived CG potentials, we analyzed the energy distributions of the sampled structures. According to the energy landscape theory (1), successful folding needs a funnel-like energy landscape. Meanwhile, the energy gap between the native state and the denatured state should be large enough to overcome energy frustrations during folding. Fig. 5, *A* and *B*, shows the energy distribution before self-learning and after three steps of self-learning, respectively, for the nativelike structures ($Q > 0.7$) and the unfolded structures ($Q < 0.4$) at $T_{CG} = 1000$ K. We fitted the energy distributions with a Gaussian function, and the peak position as a function of self-learning iteration step is plotted in Fig. 5 *C*. One can see that before the self-learning iterations, the nativelike structures and the unfolded structures have quite similar energy distributions. After three steps of self-learning iterations, the energy distribution of the nativelike structures is dramatically shifted toward the low-energy end compared to that of the unfolded structures, which implies that the self-learning iterations increase the energy bias and improve the ability to overcome energy frustrations. We also show the energy-Q plot in Fig. 5 *D*. One can see that before self-learning, the energy landscape is more random. After three steps of self-learning, the energy landscape shows a funnel-like behavior. This funnel-like energy landscape

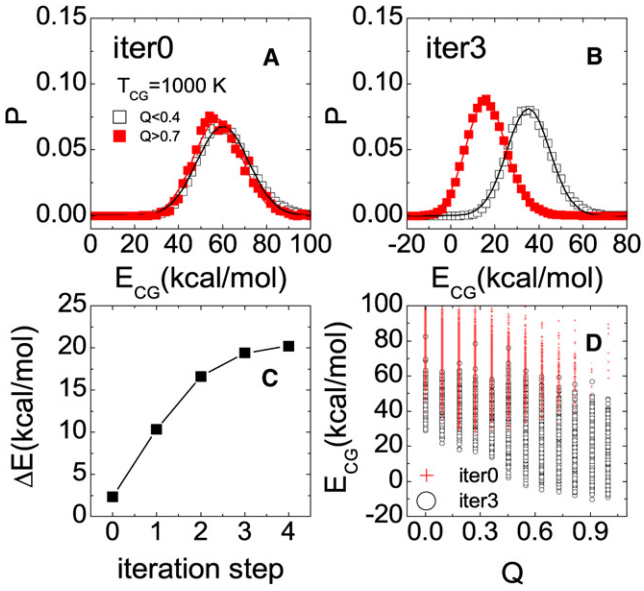

FIGURE 5 (*A* and *B*) Energy distribution for the nativelike structures ($Q > 0.7$) and the unfolded structures ($Q < 0.4$) before self-learning (*A*) and after three steps of self-learning (*B*) at $T_{CG} = 1000$ K. (*C*) Peak positions of the Gaussian fitting as a function of the iteration step. (*D*) Energy-Q plots before self-learning and after three steps of self-learning.

and the large energy bias between the native and denatured states are essential for protein folding and structure prediction, suggesting again that the SLMS method can be used to optimize CG energy functions for the purposes of protein folding and de novo structure prediction.

In the typical de novo protein structure predictions (54), CG energy functions were used to generate and filter the candidate structures at the initial stage. These candidate structures are further subjected to refinement by a high-accuracy model. Such a two-phase strategy requires that the CG energy function is accurate enough to capture at least one near-native structure during the initial CG simulations. To demonstrate the ability of the above derived CG energy function to capture the near-native structures, we picked up the low-energy CG structures sampled by CG MD simulations at 300 K after three steps of self-learning. These low-energy structures were taken from the low-energy end of the energy distribution, and they account for ~10% of the total structures. The AA details of these low-energy CG structures were then reconstructed according to the procedure described in the Methods section. The reconstructed AA structure with the smallest heavy-atom RMSD from the x-ray structure was selected and superimposed on the x-ray structure (Fig. 6, *left*). For comparison, the superimposition of the NMR structure on the x-ray structure is also shown (Fig.6, *right*). We can see that the best-fitted structure produced by the current CG energy function is quite close to the x-ray structure. The backbone (heavy-atom) RMSD is ~2.41 Å (3.16 Å), which is quite promising considering that the backbone (heavy-atom) RMSD between the x-ray

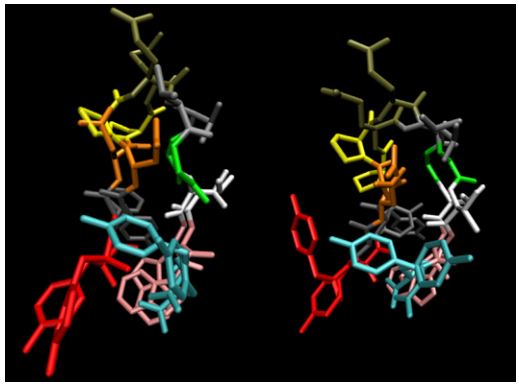

FIGURE 6 (*Left*) Superimposition (by aligning the heavy atoms) of the low-energy structure produced by CG MD (the best-fitted structure) on the x-ray structure of β-hairpin CLN025. (*Right*) Superimposition of the NMR structure on the x-ray structure.

structure and the NMR structures is ~1.75 Å (3.23 Å). Not only are the backbone atoms fitted satisfactorily, but the side-chain orientations are also roughly reproduced for most parts of the β-hairpin. There are some large discrepancies for the side-chain orientations of the terminal residues, which is possibly a result of the high flexibility of these residues in the CG simulations. The close similarity between the produced structure and the x-ray structure demonstrates that the CG energy function derived by the SLMS method is capable of capturing near-native structures during the initial stage of the de novo structure predictions.

To investigate how the CG nonlocal interactions were tuned in the learning iteration, we compared specific interactions that make the β-hairpin with those that may stabilize α-helices. For this small peptide, the most dramatic difference between the α-helices and the β-hairpin is the contact modes of the terminal residues. In the β-hairpin, Y1 forms hydrogen bonds with Y10, whereas in the α-helices, Y1 forms a hydrogen bond with E5, and Y10 forms a hydrogen bond with T6. Fig. 7 shows the CG potential between the backbone beads of the native residue pair Y1-Y10, as well as between those of the nonnative residue pairs that may form in α-helices, Y10-T6 and Y1-E5, before self-learning and after three steps of self-learning iterations. We can see that the interaction for the native contact Y1-Y10 became stronger due to the self-learning iterations. In contrast, the interactions for the nonnative residue pairs Y10-E6 and Y1-T5 either became weaker or did not change significantly. These results indicate that the self-learning iterations tend to strengthen the native interactions between Y1 and Y10, which is consistent with the principle of minimal frustration.

We then addressed the context dependence of the CG potential, namely, how CG nonlocal interactions of the same amino acid types can differ after self-learning iterations. In the x-ray structure of CLN025, there are two cases where the interacting residues in different contacts share the same amino acid types. The first is the Y-Y interaction contained in Y1-Y10 and Y2-Y10, and the other is the Y-W

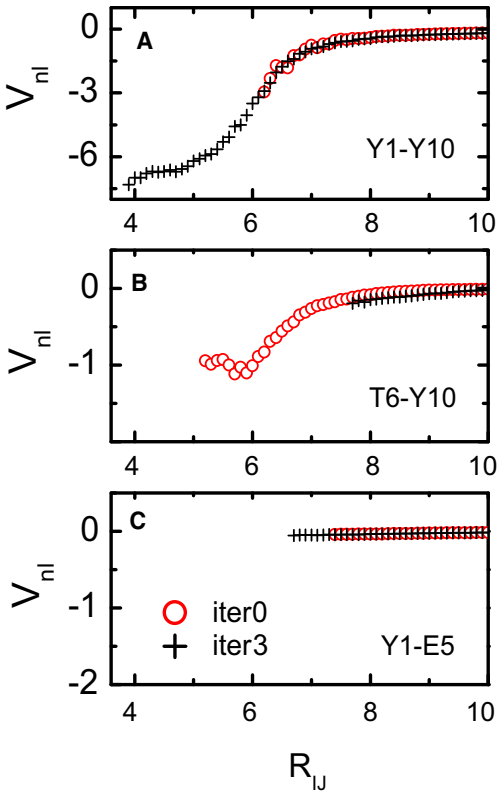

FIGURE 7 CG interactions for native and nonnative contact pairs, showing CG potentials between the backbone beads of Y1 and Y10 (*A*), T6 and Y10 (*B*), and Y1 and E5 (*C*) before self-learning (*circles*) and after three steps of self-learning (*crosses*).

interaction in Y1-W9 and Y2-W9. In the native structure, their side-chain arrangements are quite different and the residue pairs Y1-Y10 and Y2-W9 are more tightly arranged, but it should be noted that we did not use this information a priori in the SLMS. Fig. 8 shows the CG potential between the backbone beads of the residue pairs Y1-Y10, Y2-Y10, Y1-W9, and Y2-W9 before and after learning iteration. We see that the interaction between Y1 and Y10 increased more significantly than did that between Y2 and Y10 after the self-learning iterations, although they have the same side chains. In addition, the minima of the derived CG potentials for the Y1-W9 and Y2-W9 pairs correspond to those of the native structure. For example, the potential between Y2 and W9 prefers a closer contact (~4.0 Å), whereas the potential between Y1 and W9 prefers a relatively lose contact (~5.5 Å). The corresponding distances in the x-ray structure are 4.4 Å and 5.4 Å, respectively. A similar feature is found for the CG potentials of the side-chain beads of the above residues (Fig. S4). It is apparent that such correlation between pairwise interactions and contact modes in the native structure can decrease energy frustration during folding.

The above results suggest that the interaction specificity essential for correct folding to the β-hairpin can be captured to a large extent by CG potentials via the self-learning

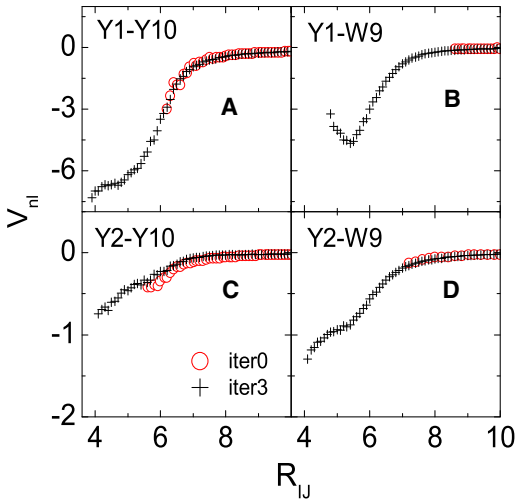

FIGURE 8 CG potentials between the backbone beads of native contact pairs Y1 and Y10 (*A*), Y1 and W9 (*B*), Y2 and Y10 (*C*), and Y2 and W9 (*D*) before self-learning (*circles*) and after three steps of self-learning (*crosses*).

process, which ensures robust folding of the β-hairpin CLN025 by the CG model. Note that the above discussions only focus on nonlocal contacts involving terminal residues to demonstrate the ability of the SLMS method to optimize the folding landscape, because the contact modes of these residues have the largest difference between β-hairpin and other structures. Actually, the pairwise interaction between the backbone beads of Y1 and Y10 is the strongest among all the pairwise interactions. Such strong interactions may result partly from the strong electrostatic interactions of the two charged termini. Although not discussed explicitly, the local and nonlocal interactions in the turn region undoubtedly also contributed to the robust folding.

It is worth mentioning that the CG force field derived by the SLMS method overstabilized the folded structures. This can be seen easily by comparing the Q distributions from the CG simulations (Fig. 4, *A–C*) and from the AA T-REMD simulations (Fig. 4 *D*). For example, the Q distribution of the CG simulations at 500 K is quite close to that of the AA T-REMD simulations at 300 K. This observation is not surprising, since we directly used the average of the decomposed AA energies as the nonlocal term of the CG potential. Theoretically, the CG potential appropriate to reproduce the results of the atomistic model is given by the many-body potential of mean force (PMF), which can be written as (11,55)

$$F(\mathbf{R}) = -k_{\mathrm{B}}T\ln\left(\int_{\mathbf{R}}\exp(-E_{AA}(\mathbf{R},\mathbf{x})/k_{\mathrm{B}}T)d\mathbf{x}\right), \quad (4)$$

where $\mathbf{R}$ stands for the CG coordinates, and $\boldsymbol{x}$ stands for the fine-grained (FG) coordinates that are averaged out during the coarse-graining. The integration is limited to the conformation space covered by $\boldsymbol{x}$ with fixed $\boldsymbol{R}$. The above many-body PMF can be further decomposed into two parts by

$F(\mathbf{R}) = < E_{AA} >_{\mathbf{R}} - TS^{FG}(\mathbf{R})$. Here, $<E_{AA}>_{\mathbf{R}}$ stands for the canonical average of the AA energies among the structures with fixed $\mathbf{R}$. The $S^{FG}(\mathbf{R})$ is the entropy arising from the FG degrees of freedom. In this work, only the first term was used as the nonlocal part of the CG potential; therefore, the contribution of the FG entropy is lost. On top, the structure mapping from the CG structures to the AA structures cannot include all the possible AA structures for each of the CG structures, which also renders incomplete the consideration of FG entropy during reweighting, as mentioned in the Methods section. Since entropy contributes to the relative stability of the denatured state, the CG potential without including the contributions from such FG entropy tends to overstabilize the folded structure and make the protein stability less sensitive to the simulation temperature, as shown in Fig. 4. These results suggest that to quantitatively calculate the thermodynamics, e.g., the folding temperature or the free-energy change during folding, we need to more thoroughly take into account the FG entropy contributions. Actually, reproducing the many-body PMF is one common challenge in most multiscale methods. Recently, Voth and co-workers developed a force-match protocol to derive the CG potential that is theoretically more rigorous and by which the many-body PMF can be described more naturally (18). It will be interesting to use the force-match protocol as the learning method of SLMS simulations in further work.

CONCLUSIONS

The SLMS method developed in our previous work was extended here by introducing the two-bead/residue coarse-graining strategy and the energy decomposition learning method to model the interaction specificity of the AA force field more accurately by the CG model. Studies for the designed β-hairpin CLN025 suggested that self-learning iterations can improve the CG potential significantly. The derived CG force field showed a funnel-like behavior and robustly folded CLN025 to its native structure, indicating that the SLMS method can be useful for studies of protein folding and structure prediction. It should be emphasized that our aim was to derive not a CG force field of general purpose, but one that was specific for the target protein. Such a tailor-made feature ensures high interaction specificity of the derived CG potential. However, care must be taken in quantitative discussions based on the SLMS used here due to the incomplete treatment of the entropy related to the FG degrees of freedom. Accurate treatment of the FG entropy calls for more delicate algorithms for structure reconstruction and learning. It is worth mentioning that recent studies have reported significant advances in the development of structure-reconstruction algorithms (31,56). For example, Liu and co-workers (56) developed a theoretically rigorous reconstruction protocol by integrating the configurational-bias Monte Carlo into the resolution exchange method. With this new method, an ensemble of AA structures for each CG structure can be reconstructed with correct weight. This theoretically rigorous method can be the starting point for future development of more efficient structure-reconstruction algorithms.

SUPPORTING MATERIAL

Four figures (Figs. S1-S4) are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)01036-2

This work was supported by a Grant-in-Aid for Scientific Research and by Research and Development of the Next-Generation Integrated Simulation of Living Matter, a part of the Development and Use of the Next-Generation Supercomputer Project of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
2. Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*. 282:740–744.
3. Jayachandran, G., V. Vishal, and V. S. Pande. 2006. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J. Chem. Phys.* 124:164902.
4. Li, W., J. Zhang, …, W. Wang. 2008. Metal-coupled folding of Cys^2His^2 zinc-finger. *J. Am. Chem. Soc*. 130:892–900.
5. Tozzini, V. 2005. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15:144–150.
6. Levitt, M., and A. Warshel. 1975. Computer simulation of protein folding. *Nature.* 253:694–698.
7. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.
8. Marrink, S. J., H. J. Risselada, …, A. H. de Vries. 2007. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B.* 111:7812–7824.
9. Okazaki, K., and S. Takada. 2008. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc. Natl. Acad. Sci. USA*. 105:11182–11187.
10. Fujitsuka, Y., S. Takada, …, P. G. Wolynes. 2004. Optimizing physical energy functions for protein folding. *Proteins.* 54:88–103.
11. Liwo, A., C. Czaplewski, …, H. Scheraga. 2001. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* 115:2323–2347.
12. Kmiecik, S., and A. Kolinski. 2008. Folding pathway of the b1 domain of protein G explored by multiscale modeling. *Biophys. J.* 94:726–736.
13. Go, N. 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:183–210.
14. Hardin, C., M. P. Eastwood, …, P. G. Wolynes. 2000. Associative memory hamiltonians for structure prediction without homology: α-helical proteins. *Proc. Natl. Acad. Sci. USA*. 97:14235–14240.
15. Tozzini, V., W. Rocchia, and J. A. McCammon. 2006. Mapping all-atom models onto one-bead coarse grained models: general properties and applications to a minimal polypeptide model. *J. Chem. Theory Comput.* 2:667–673.

16. Chu, J. W., and G. A. Voth. 2006. Coarse-grained modeling of the actin filament derived from atomistic-scale simulations. *Biophys. J.* 90:1572–1582.

17. Chu, J. W., S. Izvekov, and G. A. Voth. 2006. The multiscale challenge for biomolecular systems: coarse-grained modeling. *Mol. Simul.* 32:211–218.

18. Izvekov, S., and G. A. Voth. 2005. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B.* 109:2469–2473.

19. Trylska, J., V. Tozzini, and J. A. McCammon. 2005. Exploring global motions and correlations in the ribosome. *Biophys. J.* 89:1455–1463.

20. Chu, J. W., G. S. Ayton, …, G. A. Voth. 2007. Emerging methods for multiscale simulation of biomolecular systems. *Mol. Phys.* 105:167–175.

21. Praprotnik, M., L. D. Site, and K. Kremer. 2008. Multiscale simulation of soft matter: from scale bridging to adaptive resolution. *Annu. Rev. Phys. Chem.* 59:545–571.

22. Moritsugu, K., and J. C. Smith. 2007. Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. *Biophys. J.* 93:3460–3469.

23. Praprotnik, M., S. Matysiak, …, C. Clementi. 2007. Adaptive resolution simulation of liquid water. *J. Phys. Condens. Matter.* 19:292201.

24. Christen, M., and W. F. van Gunsteren. 2006. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *J. Chem. Phys.* 124:154106.

25. Harmandaris, V. A., N. P. Adhikari, …, K. Kremer. 2006. Hierarchical modeling of polystyrene: from atomistic to coarse-grained simulations. *Macromolecules.* 39:6708–6719.

26. Reith, D., M. Pütz, and F. Müller-Plathe. 2003. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* 24:1624–1636.

27. Müller-Plathe, F. 2002. Coarse-graining in polymer simulation: from the atomistic to the mesoscopic scale and back. *ChemPhysChem.* 3:755–769.

28. Kremer, K., and F. Muller-Plathe. 2001. Multiscale problems in polymer science: simulation approaches. *MRS Bull.* 26:205–210.

29. Li, W., and S. Takada. 2009. Self-learning multiscale simulation for achieving high accuracy and high efficiency simultaneously. *J. Chem. Phys.* 130:214108.

30. Li, W., H. Yoshii, …, S. Takada. 2010. Multiscale methods for protein folding simulations. *Methods.* 52:106–114.

31. Heath, A. P., L. E. Kavraki, and C. Clementi. 2007. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. *Proteins.* 68:646–661.

32. Thorpe, I. F., J. Zhou, and G. A. Voth. 2008. Peptide folding using multiscale coarse-grained models. *J. Phys. Chem. B.* 112:13079–13090.

33. Ayton, G. S., and G. A. Voth. 2009. Systematic multiscale simulation of membrane protein systems. *Curr. Opin. Struct. Biol.* 19:138–144.

34. Lyman, E., F. M. Ytreberg, and D. M. Zuckerman. 2006. Resolution exchange simulation. *Phys. Rev. Lett.* 96:028105.

35. Lyman, E., and D. M. Zuckerman. 2006. Resolution exchange simulation with incremental coarsening. *J. Chem. Theory Comput.* 2:656–666.

36. Liu, P., and G. A. Voth. 2007. Smart resolution replica exchange: an efficient algorithm for exploring complex energy landscapes. *J. Chem. Phys.* 126:045106.

37. Wang, Y., W. G. Noid, …, G. A. Voth. 2009. Effective force coarse-graining. *Phys. Chem. Chem. Phys.* 11:2002–2015.

38. Gohlke, H., C. Kiel, and D. A. Case. 2003. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.* 330:891–913.

39. Honda, S., T. Akiba, …, K. Harata. 2008. Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* 130:15327–15331.

40. Gront, D., S. Kmiecik, and A. Kolinski. 2007. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.* 28:1593–1597.

41. Canutescu, A. A., A. A. Shelenkov, and R. L. Dunbrack, Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* 12:2001–2014.

42. Case, D., T. Darden, …, P. Kollman. 2008. Amber 10 User's Manual. University of California, San Francisco.

43. Ferrenberg, A. M., and R. H. Swendsen. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63:1195–1198.

44. Mukherjee, A., and B. Bagchi. 2003. Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36. *J. Chem. Phys.* 118:4733–4747.

45. Kim, Y. C., and G. Hummer. 2008. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* 375:1416–1433.

46. Hornak, V., R. Abel, …, C. Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins.* 65:712–725.

47. Onufriev, A., D. Bashford, and D. A. Case. 2004. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins.* 55:383–394.

48. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.

49. Ryckaert, J., G. Ciccotti, and H. Berendsen. 1977. Numerical integration of cartesian equation of motion of a system with constraints-molecular dynamics of N-alkanes. *J. Comput. Phys.* 23:327–341.

50. Kumar, S., D. Bouzida, …, J. Rosenberg. 1992. The weighted histogram analysis method for free-energy calculations on biomolecules: 1. the method. *J. Comput. Chem.* 13:1011–1021.

51. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28.

52. Chou, P. Y., and G. D. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* 47:45–148.

53. Krishna, V., W. G. Noid, and G. A. Voth. 2009. The multiscale coarse-graining method. IV. Transferring coarse-grained potentials between temperatures. *J. Chem. Phys.* 131:024103.

54. Bradley, P., K. M. Misura, and D. Baker. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science.* 309:1868–1871.

55. Noid, W. G., J. W. Chu, …, H. C. Andersen. 2008. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* 128:244114.

56. Liu, P., Q. Shi, …, G. A. Voth. 2008. Reconstructing atomistic detail for coarse-grained models with resolution exchange. *J. Chem. Phys.* 129:114103.